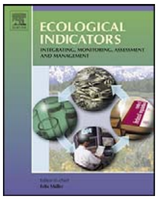
Ecological Indicators 11 (2011) 582–589

Contents lists available at ScienceDirect

Ecological Indicators

journal homepage: www.elsevier.com/locate/ecolind



Original article

Which index is the best to assess stream health?

D.A. Dos Santos*, C. Molineri, M.C. Reynaga, C. Basualdo

CONICET-Facultad de Ciencias Naturales e Instituto Miguel Lillo, Universidad Nacional de Tucumán, Miguel Lillo 205, San Miguel de Tucumán, C.P. 4000, Tucumán, Argentina

ARTICLE INFO

Article history:
Received 28 February 2010
Received in revised form 2 June 2010
Accepted 22 August 2010

Keywords:
ROC methodology
Biotic index
BMWP
EPT
Yungas
Benthic macroinvertebrates

ABSTRACT

Biotic indices are widely used in monitoring the health status of various ecosystems. The choice of the best index is generally done qualitatively depending on a variety of aspects including cost and time. ROC (Receiver Operating Characteristic) methodology constitutes a valuable tool to compare objectively the diagnostic capabilities of different tests in addition to obtain decision thresholds. In this manuscript, ROC methodology is described and implemented for the first time in the context of stream bioassessment through benthic macroinvertebrates. Cut-off values that distinguish impaired from healthy sites are suggested. A new index called IBY-4 is also developed. IBY-4 accounts for the occurrence of Megaloptera, Plecoptera, Trichoptera and Elmidae in a target site and may achieve the best general performance in the study region concerning to Andean Tropical streams.

1. Introduction

Freshwater ecosystems are one of the most endangered of the world. Moreover, the natural services they provide (mainly water) and the biodiversity they support are also threatened (UNESCO, 2009). Worldwide anthropic disturbances include channelization of stream bottom, dams, removal of riparian trees, wastewaters allocation, replacement of native forests by grasslands along the watershed and invasive species. Habitat transformation followed by biodiversity loss constitutes a consequence associated to them. Thus, the frequent supervision of the ecosystem integrity represents a priority task for water resource management. In this context, biotic indices based on aquatic macroinvertebrates have been developed as one type of diagnostic test of ecosystem integrity (for review see Bonada et al., 2006). The main premise underlying biotic indices development is that an assessment of stream integrity (and water quality) could be achieved by evaluating the community structure.

Worldwide experience has demonstrated that the most useful biological assessment methods for freshwater monitoring are based on benthic macroinvertebrates (Sivaramakrishnan, 2000). An extensive literature on this topic is available (e.g., Rosenberg and Resh, 1993; Chessman and McEvoy, 1998; Reynoldson et al., 2001; Resh, 2008). Alleged reasons are ubiquity, susceptibility to disturbances, large number of species that offers a spectrum of responses to perturbations, accessibility, inexpensive equipment for sampling, etc. (Resh, 1995). This monitoring strategy has been also incorporated to South America: e.g. Argentina (Domínguez and Fernández, 1998; Rodrigues Capítulo et al., 2001), Brasil (Junqueira and Campos, 1998), Chile (Figueroa et al., 2003, 2007), Colombia (Roldán, 1999) but with much effort devoted to adapt tolerance values or suggest the most suitable biotic index for each region (Prat et al., 2009). Fernández et al. (2002) and Von Ellenrieder (2007) represent the latest contributions studying relationships between macrobenthos and environmental variables associated to basin disturbance in the study area.

Implementation of control and protection policies should be based on indices of proven reliability. Such reliability refers to the ability of the index to detect the correct status about the health of the assessed environment and has been commonly evaluated qualitatively (e.g. Bonada et al., 2006). Nonetheless, studies that compare the performance of different biotic indices providing a statistical significance of their results are much rarer (Barbour et al., 1996; Murtaugh, 1996; Hale et al., 2004; Hale and Heltshe, 2008; Sánchez-Montoya et al., 2010).

The accuracy of a biotic index can be calculated by comparing the results of the test to the true health status of the ecosystem. True status has to be determined with reference standard procedures (chemical analyses, analysis of disturbance in the basin, etc.). To compare different biotic indices is necessary to know the following accuracy ratios: sensitivity (number of true positive predictions vs. number of actually positive cases) and specificity (number of true negative predictions vs. number of actually negative cases). The Receiver Operating Characteristic (ROC) curve is a plot of sensitivity (y coordinate) versus 1 − specificity (x coordinate). ROC curves are graphic tools especially suitable for evaluating diagnostic tests because they capture the trade-off between sen-

* Corresponding author.
E-mail addresses: dadossantos@csnat.unt.edu.ar, pseudalopex_79@yahoo.com (D.A. Dos Santos).

1470-160X/$ – see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ecolind.2010.08.004

Table 1
Two-by-two confusion matrix.

		Stream actual status	
		Perturbed (+)	Healthy (−)
Stream predicted status	Perturbed (+)	True positive (TP)	False positive (FP)
	Healthy (−)	False negative (FN)	True negative (TN)

sitivity and specificity over the range of test values (Lasko et al., 2005).

ROC curves have been applied to many disciplines, including medicine (e.g. Lusted, 1971), industrial quality control (Drury and Fox, 1975) and estuarine ecology (Hale et al., 2004; Hale and Heltshe, 2008). To our knowledge, this work represents the first contribution to the use of ROC methodology in the context of freshwater bioassessment through benthic macroinvertebrates.

The general aim of this article is to introduce basic concepts of the ROC methodology to an audience interested on freshwater biomonitoring and to emphasize its role in the appraisal of biological index performance. Specific objectives include the use of ROC curves (1) to compare the diagnostic capabilities of some widely used metrics in addition to a new index (IBY-4) applied on a large data set from Tropical Andes streams; (2) to identify thresholds of decision for those indices in order to be used in biomonitoring programs; and (3) to analyze the response of different indices to increasing levels of perturbation.

2. Materials and methods

2.1. ROC methodology

A stream is considered perturbed if it receives some anthropic impact directly on it (e.g., water chemistry or channel shifts) or on surrounding areas (e.g., riparian or watershed area denudation) to the extent of impairing the stream capability to hold a biodiversity otherwise different at the pristine condition. Basically, biomonitoring aims to determine if a given stream should be considered perturbed or not. This corresponds to a classification problem using only two classes. Formally, each instance (stream) is mapped to one element of the set {+, −} of positive (perturbed) and negative (non-perturbed) class labels (Fawcett, 2005). Classifiers are used to predict the membership of items to one of the two alternative classes. Biological metrics are classifiers that may surrogate expensive and time consuming procedures to assess the stream quality (Cullen, 1990). However, the outputs of these metrics are not single scores, they span over a range of values to which different thresholds may be applied to predict class membership. We are interested in achieving good predictions, i.e. the predicted class should agree with the actual class of stream perturbation.

Sensitivity and specificity. In dealing with predictive tasks, there are four possible outcomes: (1) *true positive*, when a perturbed stream is correctly classified; (2) *false positive*, when a healthy stream is considered an altered one; (3) *true negative*, when a preserved stream is assigned to the right class; (4) *false negative*, when a damaged stream is wrongly mapped to the non-perturbed class. The counts of correct yes-forecasts and false alarms can be arranged into a two-by-two confusion matrix (Table 1).

We will focus on two ratios, viz. the *True Positive Rate* (TPR) and the *False Positive Rate* (FPR). TPR denotes the proportion of perturbed streams correctly predicted: TPR = TP/(TP + FN); whereas FPR concerns to the proportion of negatives incorrectly classified: FPR = FP/(FP + TN). Sensitivity is equivalent to the TPR score, while specificity refers to 1 − FPR, that is the proportion of negatives correctly classified: 1 − FPR = TN/(FP + TN). Sensitivity and specificity are the basic measures of accuracy of a diagnostic test (Obuchowski, 2003); for our purposes, they describe the ability of a biological metric to correctly diagnose perturbation when perturbation is actually present and to correctly dismiss perturbation when it is truly absent.

ROC plot. Biological metrics yield a range of values rather than a dichotomous response. One strategy for obtaining binary predictions is to select a cut point and record the cases lying above and below that point. Nevertheless, the choice of a unique cut point is an arbitrary procedure that blurs the information contained in the data. As the cut point changes, specificity and sensitivity shifts (Obuchowski, 2003). A fruitful alternative is to explore the entire range of values, calculating for each possible cut point the respective sensitivity/specificity pair. The graphical display of all those pairs connected by segment lines, with sensitivity and 1 − specificity plotted on the *y* and *x* axes respectively, is known as the empirical ROC curve. Table 2 shows a workable example with the scores provided by a hypothetical metric *H* applied on 10 streams (5 perturbed and 5 non-perturbed) to illustrate how to construct the respective ROC curve (Fig. 1a). It should be considered that the true health status (gold standard) has to be fixed in a first stage of analysis and the diagnostic performance of the test has to be evaluated afterwards.

Observe that the lower the score of *H* metric, the higher the chance of predicting a positive result. A cut point at each value of *H* is established. Thus, for example, the predictions under the first criterion (i.e. perturbed if *H* < 1, otherwise non-perturbed) yield 0 for the sensitivity and 1 for the specificity, that is the point (0, 0) in the

Table 2
Hypothetical data illustrating ROC analysis. Streams actually perturbed are coded 1, otherwise they are coded 0. Values are given for an imaginary diagnostic metric called *H*. The random *H* metric is obtained via randomization of vector *H*. ROC analysis is performed below the table. For each decision threshold, 1 − specificity (1 − Spe) and sensitivity (Sen) values have been calculated. The performance of each metric can be evaluated through the respective ROC curves in Fig. 1.

	Item (status)	*H* metric Value	Random *H* Value
Raw data	Stream A (1)	1	10
	Stream B (1)	3	14
	Stream C (1)	5	1
	Stream D (1)	7	8
	Stream E (1)	10	5
	Stream F (0)	8	3
	Stream G (0)	9	12
	Stream H (0)	12	9
	Stream I (0)	14	16
	Stream J (0)	16	7
	Cut point	*H* metric (1 − Spe)/Sen	Random *H* (1 − Spe)/Sen
ROC analysis	<1	0/0	0/0
	≤1	0/0.2	0/0.2
	≤3	0/0.4	0.2/0.2
	≤5	0/0.6	0.2/0.4
	≤7	0/0.8	0.4/0.4
	≤8	0.2/0.8	0.4/0.6
	≤9	0.4/0.8	0.6/0.6
	≤10	0.4/1	0.6/0.8
	≤12	0.6/1	0.8/0.8
	≤14	0.8/1	0.8/1
	≤16	1/1	1/1

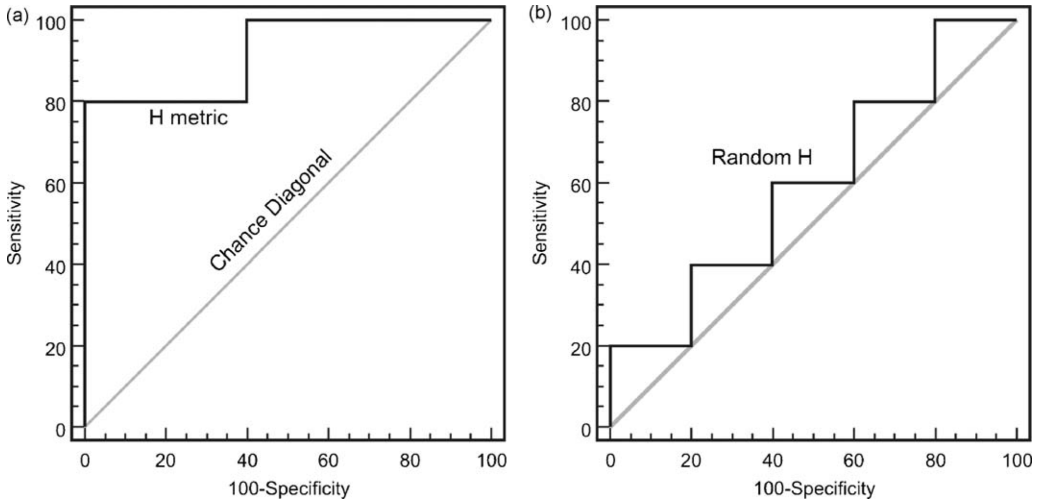


Fig. 1. Performance of hypothetical diagnostic measures visualized via ROC curves. (a) ROC curve associated to the *H* metric. (b) ROC curve associated to the randomized *H* vector. Sensitivity and specificity are expressed as percentages. Note that 100 − specificity corresponds to FPR. See Table 1 for calculations involved in the construction of these plots.

ROC space, indicating us that no positive item could be recognized, whereas all the negatives were correctly classified. A reversed result is obtained when the criterion becomes $H \leq 16$, producing here the point (1, 1) in the ROC space. This last strategy represents issuing unconditionally a positive classification where the true positive rate is maximized and all the negatives are neglected. As we move along the sorted values of the classifier variable, specificity and sensitivity experience opposite trends. Thus, the ROC graph depicts the tradeoffs between sensitivity and specificity, being an excellent tool for visualizing the performance of a diagnostic test. Since the scales of the ROC curve are the TPR and the FPR, this curve does not depend on the scale of the classification metric. This enables us to do visual comparisons among different metrics on a common set of scales (Obuchowski, 2003).

A completely randomized classifier is expected to fall along the chance diagonal associated to the ROC space. To illustrate this, see Fig. 1b where the ROC curve is displayed after random assorting of the imaginary *H* values (see Table 2). On the contrary, in case of perfect segregation between the two distributions, the ROC plot passes through the point (0, 1) indicating maximal sensitivity and specificity. Therefore, the closer the ROC plot is to the upper left corner, the higher the overall accuracy of the test (Zweig and Campbell, 1993).

Lastly, limnologists are frequently queried to give both reliable and fast judgments about the stream quality upon requests of governmental or private institutions. For those situations, it would be desirable to account for a standardized protocol, not only to sample the biological community, but also to implement more adequate metrics and to choose the less conflictive threshold decision value. These last two points can be solved through ROC analysis as we explain in the following procedures.

Optimal cut point. For the subsequent real data here analyzed, we assume that sensitivity and specificity are equally important. In this way, the threshold decision was based upon the nearest point to the upper left corner of the ROC plot. If similar performances are obtained, the cut-off with maximal sensitivity has priority. This criterion is founded on the fact that the best positioned point in the ROC space is (0, 1); so the closer a point is to this corner, the higher its performance. The nearest point to the upper left-hand corner will result in the lowest number of overall errors: FN + FP (Streiner and Cairney, 2007). Returning to the hypothetical example in Table 1 and Fig. 1, the favored decision threshold for the *H* metric should be $H \leq 7$.

AUC statistics. Given the ROC curve for a classifier, the area under the curve (AUC) measures its overall diagnostic performance. The AUC is susceptible to several interpretations (Hanley and McNeil, 1982; DeLong et al., 1988; Obuchowski, 2003), namely (1) the average value of sensitivity for all possible values of specificity, and vice versa, and (2) the probability that a randomly selected item with perturbation has an index score that indicates greater suspicion than a randomly chosen item without perturbation. Some very appealing properties of AUC are its independence on either the prevalence of perturbed items or the cut points to form the curve, and its equivalence to the Wilcoxon tests of ranks (Hanley and McNeil, 1982). Since the AUC is a portion of the area of the unit square, its value will always be between 0 and 1. However, because random guessing is associated to the chance diagonal with AUC = 0.5, no realistic classifier should have an AUC < 0.5 (Fawcett, 2005). The AUCs for the ROC curves of Fig. 1a and b are 0.92 and 0.60, respectively. Then, what does AUC = 0.92 for the *H* metric mean? If we compare the *H* values of two randomly selected streams each one with different conservation labeling, the *H* value of the perturbed stream is expected to be lower than the opposite *H* score in 92% of the times.

The 95% confidence interval for the AUC can be calculated to test random deviation from the null hypothesis of AUC = 0.5. If the 0.5 value lies outside the confidence interval, then there is evidence that the metric is able to distinguish between the two groups of streams. When different diagnostic metrics are applied on the same streams, their ROC curves may be also tested for the statistical significance of the difference among their AUCs scores (DeLong et al., 1988). See Hanley and McNeil (1982) and Zweig and Campbell (1993) for details of calculations. ROC analyses and graphics were performed using MedCalc for Windows, version 9.6.4.0 (MedCalc Software, Mariakerke, Belgium, available at http://www.medcalc.be) and R (R Development Core Team, 2009).

2.2. Study sites, establishing scenarios of perturbation and stream quality indices

Southern Andean Yungas (Olson et al., 2001) is a narrow strip of mountain rain forest, that extends on the eastern slope of the Andes, stretching from southern Bolivia (23°S) to northwestern Argentine (29°S). The region is characterized by a humid climate, with rainfall exceeding 1500 mm per year concentrated in the summer period (November–March). We have studied 95 streams (corresponding

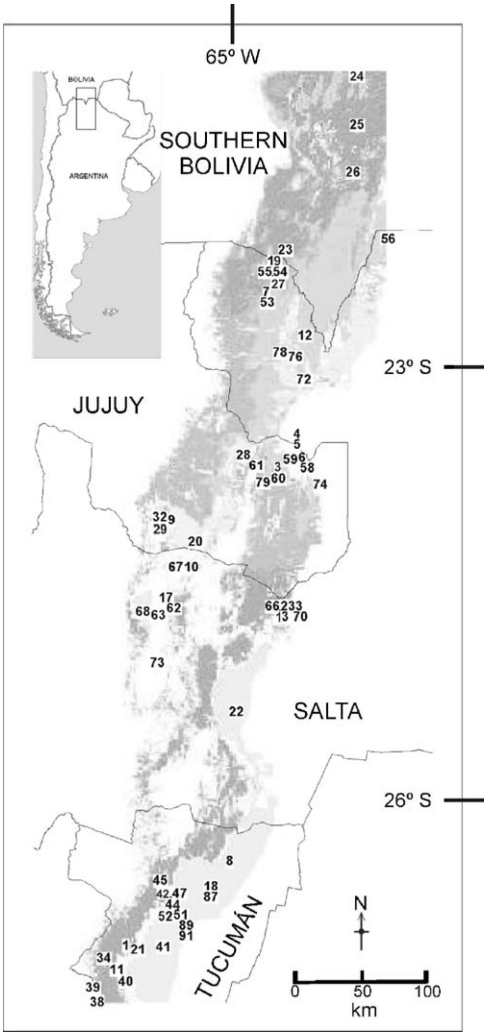


Fig. 2. Study area with sampling points projected on it. Grey tones correspond to vegetation formations at Yungas Rainforest.

to 95 sampling sites) widely scattered over the region that have been considered as belonging to a single lotic typology (Fig. 2). Sampled streams show steeped slopes, moderately to fast current velocity, well oxygenated and temperate waters, with streambeds dominated by boulders and cobbles. Presence/absence data about benthic invertebrates were obtained via qualitative (kicknet and light trap) and quantitative (Surber 0.3 m × 0.3 m, 300 μ) sampling methods with similar search efforts. We have identified 171 taxa to species/morpho-species level (Ephemeroptera, Trichoptera, Plecoptera, Megaloptera and the family Elmidae of Coleoptera) and higher levels for the remaining taxa.

The true health status of streams was *a priori* determined in function of disturbance factors described below. Then, some biotic indices based on benthic assemblages were calculated. Finally, we applied ROC analysis over the metrics and obtained a graphical display of their performance and an optimal cut point to distinguish healthy from impaired streams. The above steps can be articulated into a single analytic pathway (Fig. 3).

Perturbation was assessed based on the occurrence (or not) of the following disturbance factors (up to 1 km upward the main flow direction): logging (restricts to the riparian trees), cattle grassing, physical alteration of the stream bottom (canalization, vehicular traffic passing over the streambed, and commercialization of pebbles, cobbles and sand), recreational or domestic use of the stream by local population, replacement of native forests by crops, and settlement of factories that discharge effluents on the running waters. Evidence was provided by field notes, aerial image interpretation and literature (Fernández et al., 2001; Von Ellenrieder, 2007). The variable here considered are inside the pool of criteria commonly used to define reference sites (e.g., Sánchez-Montoya et al., 2009). Four different scenarios of perturbation have been evaluated. In the first scenario, the true health status of a given stream is classified as positive (perturbed) if only one disturbance factor has been detected, in the second scenario a stream is classified as positive if two factors have been detected, and so on. Note that this classification depends only on the number of impacts disregarding their intensity.

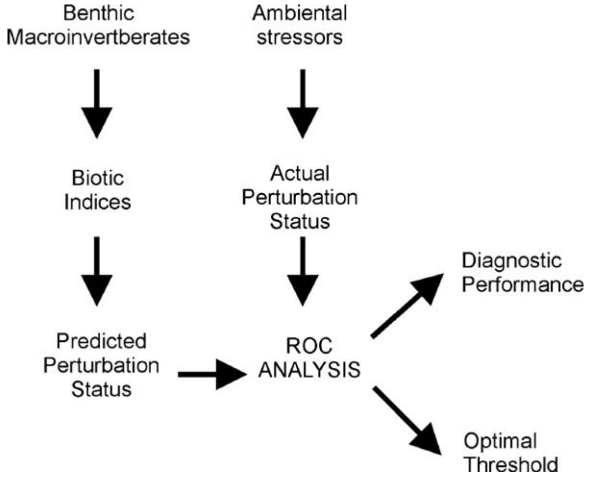


Fig. 3. Analytic pathway of ROC analysis applied on stream quality inference. Two independent channels can be recognized, one determines the actual conservation level studying environmental descriptors and the other aims to predict that status via a biotic index. ROC analysis assesses the diagnostic performance and helps to define a cut-off to decide whether a given stream should be classified as impaired or healthy.

Six biotic indices (Table 3) were calculated. From that list, IBY-4 (*Yungas Biotic Index based on 4 taxa*) is firstly proposed here. It accounts for the occurrence of Elmidae, Plecoptera, Trichoptera and Megaloptera. IBY-4 is independent of the richness and abundance associated to each of these taxa. The domain of IBY-4 is thus constrained to a set of five discrete states {0, 1, 2, 3, 4}. IBY-4 = 0 means that none of the four taxa has been detected along the sampling site, IBY-4 = 1 means that only one of the four taxa has been recorded, and so on. These four taxa have been selected because: (1) they have shown good responses on separate ROC analyses performed by the authors over the totality of available taxa; (2) they are conspicuous elements of the Andean stream communities; (3) they represent easily recognizable taxonomic levels, therefore they are a convenient tool to assess stream quality in the field even for parataxonomists.

3. Results

Thirty one streams did not exhibit any recognizable disturbance factors acting on them as they are mainly associated to protected areas and were treated as reference sites. The frequency of streams with 1, 2, 3 or 4 disturbance factors was 31, 10, 10 and 13, respectively. ROC curves enabled us to evaluate different metrics and their responses for each scenario of perturbation (Figs. 4 and 5). The first scenario was significantly solved by the ASPT metric. The second scenario was discriminated by ASPT and IBY-4. The third scenario was recognized through all indices except Richness, and the last scenario was discriminated through all the metrics. Table 4 displays the basics statistics of the ROC analyses, including both AUC estimation (the general performance of

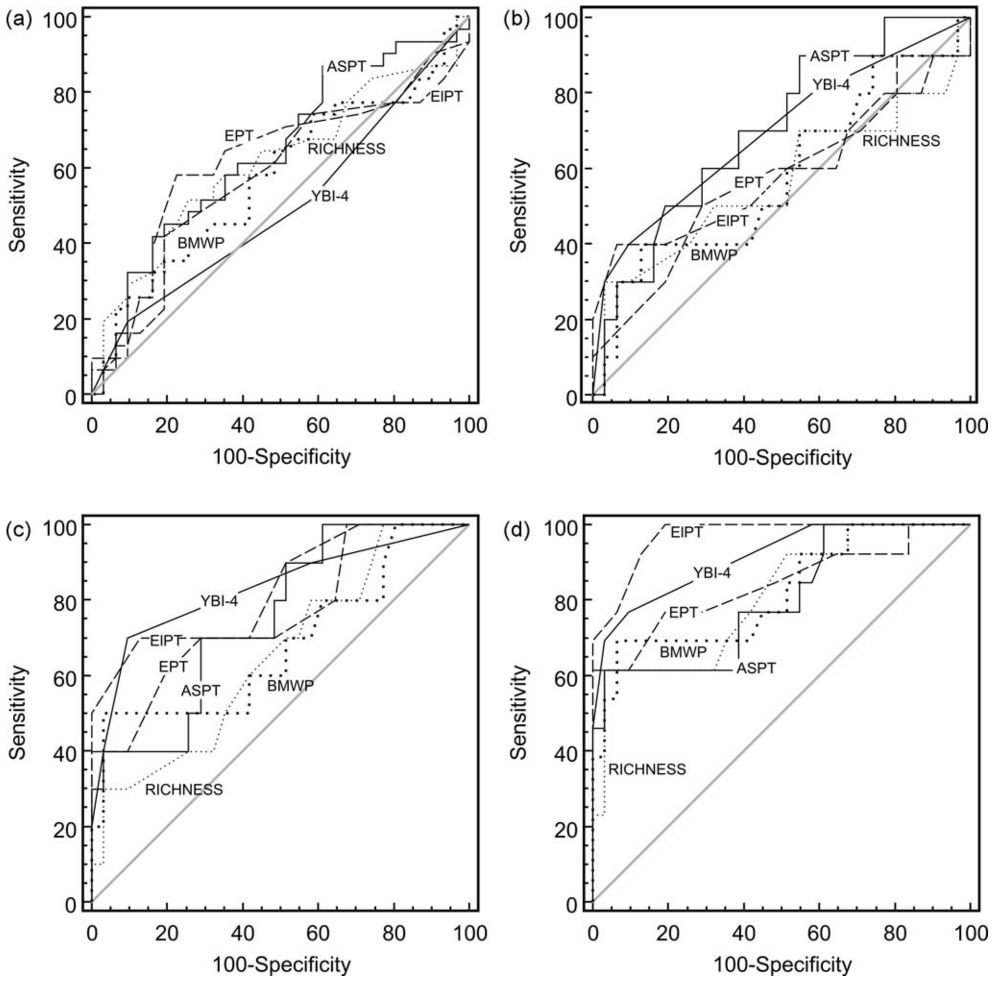


Fig. 4. Diagnostic performance of biotic indices in different scenarios of perturbation. (a–d) ROC curves. Diagonal chance has been drawn. The response variable (perturbation) corresponds to the presence/absence of one (a), two (b), three (c), or four (d) disturbance factors. Sensitivity/specificity values are expressed in percentages.

the test) and information about cut points. The mean AUC scores may achieve an overall hierarchy of performance, where IBY-4 > ASPT > ElPT > EPT > BMWP > Richness. In general terms, all the metrics increased their performance as the lower bound of perturbation also increased from 1 to 4 (Fig. 5).

The performance of the optimal cut points for the different metrics can also be assessed. In Fig. 6 each point on the ROC space represents a sensitivity/specificity pair corresponding to a particular decision threshold. Different perturbation scenarios (closed by parentheses) present some variations but in general ElPT and IBY-4 outperform the remaining indices. ROC curve methodology also permitted to estimate the cut-off values for each metric. We found that the average decision thresholds to ascertain perturbation are: IBY-4 ≤ 2, ASPT ≤ 6, ElPT ≤ 4, EPT ≤ 8, BMWP ≤ 66 and Richness ≤ 14.

Despite the fact that ElPT and EPT are highly related indices, the ROC curves of the former consistently dominated on almost the entire specificity domain of the latter (Fig. 4). Fig. 7a displays the separate performance for each taxa involved in these biotic indices. Although Ephemeroptera seems not to be completely satisfactory, a different picture is achieved when its main families are analyzed separately (Fig. 7b). Baetidae and Caenidae do not differ from a random guessing of stream quality, while Leptohyphidae and Leptophlebiidae are better discriminators.

Table 3
Metrics applied on real data to show ROC analysis.

Metrics	Description	Observation
Richness	Number of different kinds of organisms recorded at a given site.	The simplest estimation of biodiversity.
BMWP (*Biological Monitoring Working Party*)	Sum of the tolerance scores of all families recognized in the sample (Armitage et al., 1983).	We used scores adapted for the region by Domínguez and Fernández (1998).
ASPT (*Average Score Per Taxon*)	Total BMWP divided by the number of scoring taxa (Walley and Hawkes, 1996).	Dependence on sample size is removed.
EPT (*Ephemeroptera–Plecoptera–Trichoptera*)	Number of species/morpho-species within these orders (Klemm et al., 1990).	Pollution sensitive taxa pooled into a single measure.
ElPT (*Elmidae–Plecoptera–Trichoptera*)	Number of species/morpho-species within these taxa.	Von Ellenrieder (2007) prefers this index over EPT for the region of interest.
IBY-4 (*Yungas Biotic Index based on 4 taxa*)	It accounts for the occurrence of Elmidae, Plecoptera, Trichoptera and Megaloptera.	Contribution of this manuscript. See text for explanation.

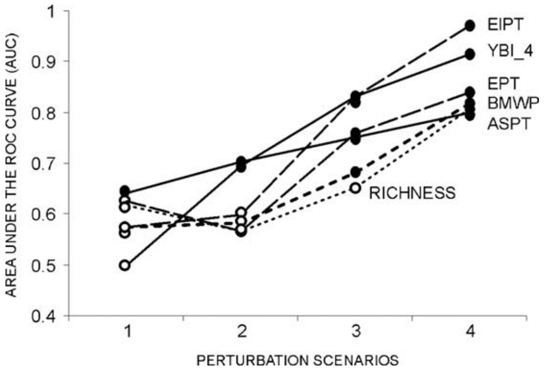


Fig. 5. AUC values under the three different scenarios of perturbation. Solid circles: AUC significantly higher than random expectation. Empty circles: the null hypothesis (AUC = 0.5) could not be rejected at a significance level of 0.05.

4. Discussion

4.1. Test comparisons

International standardization of methods is increasingly required around the world, for example in Europe by the Water Framework Directive (Bloch, 1999), and ROC curve methodology greatly outperforms previous methods used for the comparison of test efficiency (i.e. accuracy). The percentage of correct diagnoses in the entire sample can be simply computed but has several limitations (Obuchowski, 2003): (1) its magnitude varies as the prevalence of disturbed sites varies in the sample, (2) it is calculated on the basis of only one cut point and (3) false-positive and false-negative results are treated as if they are equally undesirable. ROC curve analysis, instead, combines sensitivity and specificity without creating a dependence on the prevalence of impacted sites. Furthermore, the ROC plot displays all possible cut points (or sensitivity/specificity pairs) enabling a direct visual comparison of many tests on a common set of scales.

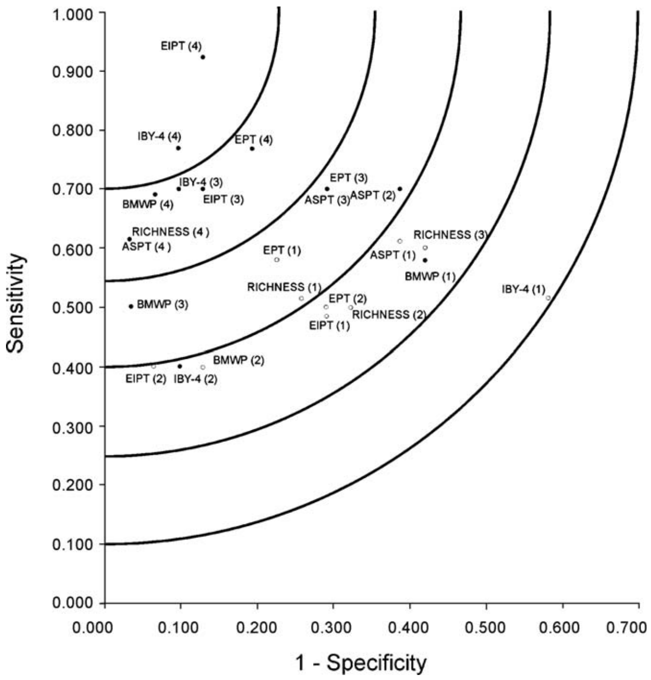


Fig. 6. Single performance of the optimal cut points for the different metrics. Each point on the ROC space represents a sensitivity/specificity pair corresponding to a particular decision threshold. Classifiers approaching to the upper left corner should be preferred. The classes of perturbation scenario are closed by parentheses. Circular tracks indicate sensitivity/specificity pairs that are equidistant to the point (0, 1). In general, metrics improve their performance (i.e., they are northwest) when the minimal level of perturbation considered raises up. ElPT and IBY-4 outperform the rest of indices.

Barbour et al. (1996) developed a framework to assess discrimination between impaired and unimpaired sites by stream biological metrics. These authors consider accuracy according to the degree of interquartile overlap in paired box-and-whisker plots. This analytic strategy is rather similar to the underlying logic of ROC analysis (i.e. to compare two series of test scores based on the ordinal arrangement of metric values). However, Barbour et al.'s approach does not provide (1) guidelines to differentiate the discriminatory ability of

Table 4
Summary statistics for the ROC curve analyses. Biotic indices are the classifiers and scenarios of perturbation correspond to the actual classification of streams. Scenarios 1, 2, 3 and 4 are based on the corresponding number of disturbance factors acting on the streams. For each index the last column indicates the best cut-off (i.e., with the maximal sum of ranks) throughout the scenarios. Scenarios with non-significant AUC values show empty entries for cut-off, sensitivity and specificity. Sen = sensitivity, 1 − Spe = 1 − specificity.

Metric	Scenario	AUC	Cut-off	Sensitivity	Specificity	Mean AUC	Best cut-off
IBY-4	1	0.499	–	–	–	0.735	≤2
	2	0.695	≤2	0.400	0.903		
	3	0.831	≤2	0.700	0.903		
	4	0.913	≤2	0.769	0.903		
ASPT	1	0.640	≤6.210	0.613	0.613	0.724	≤6
	2	0.702	≤6.187	0.700	0.613		
	3	0.752	≤6	0.700	0.710		
	4	0.800	≤5	0.615	0.968		
ElPT	1	0.573	–	–	–	0.707	≤4
	2	0.598	–	–	–		
	3	0.832	≤4	0.700	0.871		
	4	0.970	≤4	0.923	0.871		
EPT	1	0.624	–	–	–	0.635	≤8
	2	0.566	–	–	–		
	3	0.758	≤8	0.700	0.710		
	4	0.840	≤7	0.769	0.807		
BMWP	1	0.574	–	–	–	0.627	≤66
	2	0.581	–	–	–		
	3	0.681	≤51	0.500	0.968		
	4	0.818	≤66	0.692	0.936		
Richness	1	0.616	–	–	–	0.602	≤14
	2	0.565	–	–	–		
	3	0.652	–	–	–		
	4	0.808	≤14	0.615	0.968		

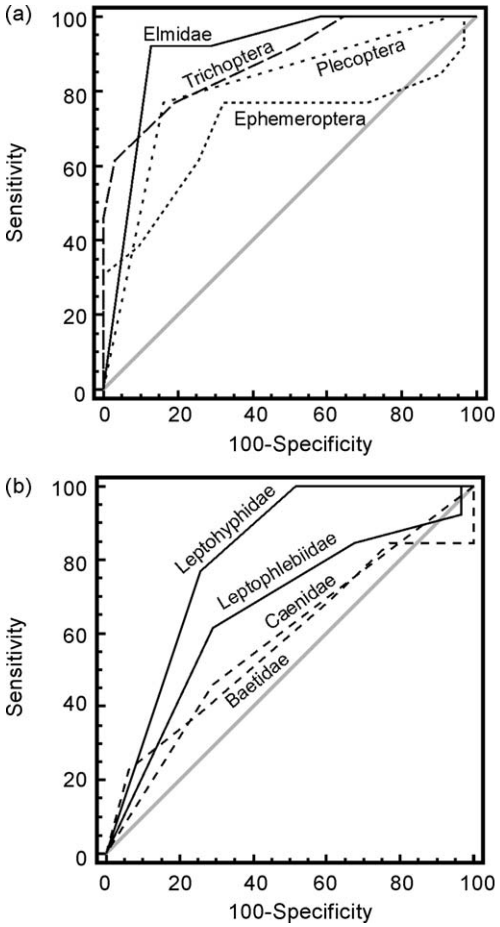


Fig. 7. Differential diagnostic performance among taxa involved in the calculation of ElPT and EPT metrics under the fourth scenario (most perturbed). (a) Individual ROC curves. (b) ROC curves associated to the main families of Ephemeroptera.

metrics belonging to the same category of accuracy, (2) thresholds separating healthy from poor streams based on an optimality criterion.

Another strategy for carrying out comparisons among methods consists of enumerating a list of attributes (or ideal criteria) that an index should meet (Bonada et al., 2006). Preferred indices are those that fulfill the maximum number of these criteria. This kind of qualitative evaluation is very different to the quantitative one discussed here, and is useful to determine the more suitable method prior to the development of the monitoring effort or when a gold standard cannot be achieved by the available data for the study region.

Our findings show that ASPT outperformed all other metrics in the first scenario (pristine vs. rivers with only 1 impact factor). So, this metric is a useful tool for monitoring areas impacted by diffuse factors (patchy landscapes with native forest, pastures and crops). The use of any other of the metrics discussed here would fail in detecting this impact on the aquatic community.

As expected, all the metrics increased their performance with increasing perturbation. Noticeably, the AUC for Richness and other indices linked to richness (ElPT, EPT, BMWP) decreased (i.e. indices show a lower discriminatory resolution) throughout the second scenario (Fig. 5). This pattern accommodates to the intermediate disturb hypothesis (Connell, 1978). Intermediate-level factors are known to increase diversity in many cases and the second scenario subsumes into this case. Moreover, the second scenario includes rivers with some organic enrichment and/or other factors creating spatial heterogeneity.

As proposed by Von Ellenrieder (2007) ElPT outperformed EPT which is a phenomenon already suggested by Fossati et al. (2001) who highlighted Elmidae strong responses to sediments loads. In regard to EPT, the exclusion of some tolerant families of Ephemeroptera (Baetidae, Caenidae) from EPT calculation improved its effectiveness. For example Baptista et al. (2007) discussed the metric Baetidae/Ephemeroptera that is expected to increase in the presence of perturbation. This consideration was also pointed out by Domínguez and Fernández (1998) in adapting BMWP scores to the region (giving low sensitivity value assigned to Baetidae). Thus, the conclusions about performance of Ephemeroptera may be different depending on the taxonomical level evaluated.

BMWP performed relatively well in the worst scenarios (with 3 and 4 impacts), but failed to determine low affected systems (with 1 or 2 impacts). This failure may be associated to the general score allocation for each taxon which has to be adjusted to the region as the knowledge about fauna pollution tolerance increases (Prat et al., 2009). The *ad hoc* index IBY-4 showed high AUCs throughout the different scenarios. Its extremely simple calculation only accounts for the recognition of four higher taxa (Elmidae, Trichoptera, Plecoptera and Megaloptera) with relatively large body size and easily identifiable. Presence/absence of these taxa is solely required to obtain IBY-4 score, so morpho-species or species are not necessary to be identified. These aspects make IBY-4 a powerful tool for biomonitoring by non-taxonomists, and especially suitable for local people to carry out a rapid assessment of the ecological stream quality.

4.2. Cut-off values

Classic ROC methodology is concerned with variables of dichotomous response, so cut-off values chosen via this tool separate two opposite categories: healthy versus impaired. However, some metrics like BMWP are commonly used to discriminate more than two categories of stream status conservation (e.g., pristine, good, regular, bad). Our cut-off value for this metric (=66) distinguishes pristine + good from regular + bad streams. Nonetheless, multi-class ROC analysis can be used to discriminate outcomes associated to multiple underlying categories (Li and Fine, 2008).

An advantage of using ROC analysis in bioindication refers to the estimation of cut-off values. Traditionally, cut-off values are estimated through the comparison of metric values gathered from impacted and polluted sites, in a rather circular reasoning. The cut-off value for BMWP ($\leq$66) is much higher than that proposed originally for the region ($\leq$40, Domínguez and Fernández, 1998). In this particular case, the difference is due to a better knowledge of the region (Fernández et al., 2006) with a larger sampling of sites (95 vs. 17) and taxa (171 vs. 34) and the inclusion of streams holding a biodiversity higher than those studied by Domínguez and Fernández (1998).

The other commonly used metrics (ASPT $\leq$ 6, ElPT $\leq$ 4, EPT $\leq$ 8, Richness $\leq$ 14) and the new IBY-4 ($\leq$2), did not have a proposed cut-off value, so the outcome obtained here are the first proposals.

4.3. Future research

The success in implementing biotic indices is subordinated to the accuracy of the sensitivity scores assigned to individual taxa. Thus, for example, Figueroa et al. (2007) working on Chilean Mediterranean rivers propose to tune the tolerance values of freshwater benthic macroinvertebrates for that region so that discriminatory capabilities of biotic indices could be enhanced. A very fertile field of research associated to ROC curves, not explored here, is on the allocation of BMWP scores to taxa in an objective way. For example in Fig. 7a, we can see that Elmidae clearly outperforms Perlidae (the single family of Plecoptera in the region) in discriminating polluted sites, but Perlidae has the maximum BMWP score (10) whereas Elmidae shows a low score (5) (Domínguez and Fernández, 1998). Our results suggest that Elmidae should have a much higher value in the studied area.

We have considered different scenarios of perturbation throughout the paper which ranged from pristine to highly perturbed ones based on the number of occurring disturbance factors. However, study sites are exposed to different levels for each disturbance factor, so that weighting the impact of those factors would be desirable. There is no gold standard to classify a given stream as healthy or non-healthy; on the contrary, there is a continuous stressor gradient in river ecosystems. We consider the ROC fuzzy methodology (Castanho et al., 2007) a valuable tool to address this issue as this approach considers outcome variables not only as dichotomous but also as continuous ones.

Acknowledgments

We especially thank to the research group Instituto de Biodiversidad Neotropical (Facultad de Cs. Naturales e IML, UNT), Priscilla Powell and Natalia von Ellenrieder for providing the data analyzed here. We also thank Eduardo Domínguez, Núria Bonada, Narcís Prat and Hugo Fernández for manuscript revision and useful comments. Ana Soledad Pérez Toranzos revised the English of this manuscript. This paper has received financial support from CONICET, ANPCYT-PICT 01-528 and PIP 1484. Two anonymous referees and editors made valuable suggestions.

References

Armitage, P.D., Moss, D., Wright, J.F., Furse, M.T., 1983. The performance of a new biological water quality score system based on macroinvertebrates over a wide range of unpolluted running water sites. Water Res. 17, 333–347.

Barbour, M.T., Gerritsen, J., Griffith, G.E., Frydenborg, R., McCarron, E., White, J.S., Bastian, M.L., 1996. A framework for biological criteria for Florida streams using benthic macroinvertebrates. J. N. Am. Benthol. Soc. 15, 185–211.

Baptista, D., Buss, D., Egler, M., Giovanelli, A., Silveira, M., Nessimian, J., 2007. A multimetric index based on benthic macroinvertebrates for evaluation of Atlantic Forest streams at Rio de Janeiro State, Brazil. Hydrobiologia 575, 83–94.

Bloch, H., 1999. European Water Policy Facing the New Millennium: The Water Framework Directive, in Assessing the ecological Integrity of Running Waters. Viena, Austria, pp. 9–11.

Bonada, N., Prat, N., Resh, V.H., Statzner, B., 2006. Developments in aquatic insect biomonitoring: a comparative analysis of recent approaches. Annu. Rev. Entomol. 52, 495–523.

Castanho, M.J., Barros, L.C., Yamakami, A., Vendite, L.L., 2007. Fuzzy receiver operating characteristic curve: an option to evaluate diagnostic tests. IEEE T. Inf. Technol. B 11, 244–250.

Chessman, B.C., McEvoy, P.K., 1998. Towards diagnostic biotic indices for river macroinvertebrate. Hydrobiologia 364, 169–182.

Connell, J., 1978. Diversity in tropical rain forests and coral reefs. Science 199, 1302–1310.

Cullen, P., 1990. Biomonitoring and environmental management. Environ. Monit. Assess. 14, 107–114.

DeLong, E.R., DeLong, D., Clarke-Pearson, D.L., 1988. Comparing the area under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 44, 837–845.

Domínguez, E., Fernández, H.R., 1998. Calidad de los ríos de la cuenca del Salí (Tucumán, Argentina) medida por un índice biótico. Conservación de la Naturaleza, N° 12. Fundación Miguel Lillo, Tucumán.

Drury, C.G., Fox, J.G., 1975. Human Reliability in Quality Control. Halsted, New York.

Fawcett, F., 2005. An introduction to ROC analysis. Pattern Recogn. Lett. 27, 861–874.

Fernández, H.R., Romero, F., Peralta, M., Grosso, L., 2001. La diversidad del zoobentos en ríos de montaña del noroeste de Argentina: comparación entre seis ríos. Ecología Austral 11, 9–16.

Fernández, H.R., Romero, F., Vece, M.B., Manzo, V., Nieto, C., Orce, M., 2002. Evaluación de tres índices bióticos en un río subtropical de montaña (Tucumán-Argentina). Limnetica 21, 1–13.

Fernández, H.R., Domínguez, E., Romero, F., Cuezzo, M.G., 2006. La calidad del agua y la bioindicación en los ríos de montaña del Noroeste Argentino. Serie Conservación de la Naturaleza 16. Fundación Miguel Lillo, Tucumán.

Figueroa, R., Valdovinos, C., Araya, E., Parra, O., 2003. Macroinvertebrados bentónicos como indicadores de calidad de agua de ríos del sur de Chile. Rev. Chil. Hist. Nat. 76, 275–285.

Figueroa, R., Palma, A., Ruiz, V., Niel, X., 2007. Análisis comparativo de índices bióticos utilizados en la evaluación de la calidad de las aguas de un río mediterráneo de Chile: río Chillán, VIII Región. Rev. Chil. Hist. Nat. 80, 225–242.

Fossati, O., Wasson, J., Héry, C., Salinas, G., Marín, R., 2001. Impact of sediment releases on water chemistry and macroinvertebrate communities in clear water Andean streams (Bolivia). Arch. Hydrobiol. 151, 33–50.

Hale, S.S., Paul, J.F., Heltshe, J.F., 2004. Watershed landscape indicators of estuarine benthic condition. Estuar. Coast. 27, 283–295.

Hale, S.S., Heltshe, J.F., 2008. Signals from the benthos: development and evaluation of a benthic index for the nearshore Gulf of Maine. Ecol. Indic. 8, 338–350.

Hanley, J.A., McNeil, B.J., 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143, 29–36.

Junqueira, V.M., Campos, S.C.M., 1998. Adaptation of the "BMWP" method for water quality evaluation to Rio das Velhas watershed (Minas Gerais, Brazil). Acta Limnol. Bras. 10, 125–135.

Klemm, D.J., Lewis, P.A., Fulk, F., Lazorchak, J.M., 1990. Macroinvertebrate Field and Laboratory Methods for Evaluating the Biological Integrity of Surface Waters. EPA/600/4-90/030. U.S. Environmental Protection Agency. Environmental Monitoring Systems Laboratory, Cincinnati, Ohio 45268.

Lasko, T.A., Bhagwat, J.G., Zou, K.H., Ohno-Machado, L., 2005. The use of receiver operating characteristic curves in biomedical informatics. J. Biomed. Inform. 38, 404–415.

Li, J., Fine, J.P., 2008. ROC analysis with multiple classes and multiple tests: methodology and its application in microarray studies. Biostatistics 9, 566–576.

Lusted, L.B., 1971. Signal detectability and medical decision making. Science 171, 1217–1219.

Murtaugh, P.A., 1996. The statistical evaluation of ecological indicators. Ecol. Appl. 6, 132–139.

Obuchowski, N.A., 2003. Receiver operating characteristic curves and their use in radiology. Radiology 229, 3–8.

Olson, D., Dinerstein, E., Hedao, P., Walters, S., Allnutt, P., Loucks, C., Kura, Y., Kassem, K., Webster, A., Bookbinder, M., 2001. Terrestrial Ecoregions of the Neotropical Realm (map). Conserv. Sci. Program, WWF-US, DC.

Prat, N., Rios, B., Acosta, R., Rieradevall, M., 2009. Los macroinvertebrados como indicadores de la calidad de las aguas. In: Domínguez, E., Fernández, H.R. (Eds.), Macroinvertebrados bentónicos sudamericanos. Sistemática y biología. Fundación Miguel Lillo, Tucumán, pp. 631–654.

R Development Core Team, 2009. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org.

Resh, V.H., 1995. Freshwater benthic macroinvertebrates and rapid assessment procedures for water quality monitoring in developing and newly industrialized countries. In: Davis, W.S., Simon, T.P. (Eds.), Biological Assessment and Criteria. Tools for Water Resource Planning and Decision Making. CRC Press, U.S.A., pp. 167–177.

Resh, V.H., 2008. Which group is best? Attributes of different biological assemblages used in freshwater biomonitoring programs. Environ. Monit. Assess. 138, 131–138.

Reynoldson, T.B., Rosenberg, D.M., Resh, V.H., 2001. Comparison of models predicting invertebrate assemblages for monitoring in the Fraser River catchment, British Columbia. Can. J. Fish. Aquat. Sci. 58, 1395–1410.

Rodrigues Capítulo, A., Tangorra, M., Ocón, C., 2001. Use of benthic macroinvertebrates to assess the biological status of Pampean streams in Argentina. Aquat. Ecol. 35, 109–119.

Roldán, G., 1999. Los macroinvertebrados y su valor como indicadores de la calidad del agua. Rev. Acad. Colomb. Cienc. 23, 375–387.

Rosenberg, D.M., Resh, V.H., 1993. Freshwater Biomonitoring and Benthic Macroinvertebrates. Chapman & Hall, New York.

Sánchez-Montoya, M.M., Vidal-Abarca, M.R., Puntí, T., Poquet, J.M., Prat, N., Rieradevall, M., Alba-Tercedor, J., Zamora-Muñoz, C., Toro, M., Robles, S., Álvarez, M., Suárez, M.L., 2009. Defining criteria to select reference sites in Mediterranean streams. Hydrobiologia 619, 39–54.

Sánchez-Montoya, M.M., Vidal-Abarca, M.R., Suárez, M.L., 2010. Comparing the sensitivity of diverse macroinvertebrate metrics to a multiple stressor gradient in Mediterranean streams and its influence on the assessment of ecological status. Ecol. Indic. 10, 896–904.

Sivaramakrishnan, K.G., 2000. A refined rapid bioassessment protocol for benthic macroinvertebrates for use in peninsular Indian streams and rivers. In: Ramachandra, T.V., Rajasekara Murthy, Co, Ahalya, N. (Eds.), Proceedings of Lake 2000—Symposium on Restoration of Lakes and Wetlands. Center for Ecological Sciences, IISc, Bangalore, pp. 302–314.

Streiner, D.L., Cairney, J., 2007. What's under the ROC? An introduction to receiver operating characteristics curves. Can. J. Psychiat. 52, 121–128.

UNESCO, 2009. Water in a changing world. In: The United Nations World Water Development Report 3. UNESCO Publishing.

Von Ellenrieder, N., 2007. Composition and structure of aquatic insect assemblages of Yungas mountain cloud forest streams in NW Argentina. Rev. Soc. Entomol. Argent. 66, 57–76.

Walley, W.J., Hawkes, H.A., 1996. A computer-based reappraisal of the Biological Monitoring Working Party scores using data from the 1990 River Quality Survey of England and Wales. Water Res. 30, 2086–2094.

Zweig, M.H., Campbell, G., 1993. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clin. Chem. 39, 561–577.